Data collection of cue phrases in the DEAL domain

Anna Hjalmarsson
Centre for Speech Technology
KTH
Stockholm, Sweden
annah@speech.kth.se

Jenny Klarenfjord
Centre for Speech Technology
KTH
Stockholm, Sweden
jennykl@kth.se

Abstract

This paper describes a data collection which is the foundation for an effort towards more human-like language generation in DEAL, a spoken dialogue system developed at KTH. One particular aim was to distinguish different types of *cue phrases* used in the DEAL domain. Cue phrases are linguistic devices used to signal how new dialogue contributions relate to previous discourse. Two annotators labelled cue phrases in the collected corpus with high inter-annotator agreement (kappa coefficient 0.82). A subsequent listening test on the perception of the responsive cue phrase "ja" (Eng: yes) revealed a low agreement when categorising responsives into four discourse-pragmatic categories. A slightly higher agreement on stimuli with context suggests that humans perceive the discourse-pragmatic function of "ja" based on contextual features rather than the acoustic realization of the word.

1 Introduction

In this paper we report on a data collection of human-human dialogue aiming at extending the knowledge of human-human interaction and in particular to distinguish different types of cue phrases used in the DEAL domain. DEAL is a spoken dialogue system for conversational training for second language learners of Swedish. The DEAL objectives are to build a system which is fun, human-like, and engaging to talk to, and which gives second language learners of Swedish conversation training (as described in Hjalmarsson et al., 2007). DEAL sets the scene of a flea market where a talking animated agent is the owner of a shop where used objects are sold. The student is given a mission: to buy items from the shopkeeper at the best possible price by bargaining. In spontaneous conversation humans produce speech incrementally and on-line as the dialogue progresses using information from several different sources in parallel. When starting to speak, we typically do not have a complete plan of how to say something or even what to say. Yet, we manage to rapidly integrate information from different sources in parallel and simultaneously plan and realize new dialogue contributions. In order to generate output incrementally in DEAL we need extended knowledge on how to signal relations between different segments of speech.

2 The DEAL corpus collection

The dialogue data recorded was informal, human-human, face-to-face conversation. The task and the recording environment were set up to mimic the DEAL domain and role-play. The data collection was made with 6 subjects (4 male and 2 female), 2 posing as shopkeepers and 4 as potential buyers. Each customer interacted with the same shopkeeper twice, in two different scenarios. Each dialogue was about 15 minutes long, so about 2 hours of speech were collected altogether. The shopkeepers used an average of 13.4 words per speaker turn while the buyers' turns were generally shorter, 8.5 words per turn (in this paper turn always refers to speaker turns). In total 16357 words were collected. All dialogues were first transcribed orthographically including non-lexical entities such as laughter and hawks. Filled pauses, repetitions, corrections and restarts were also labelled manually. The labelling of cue phrases included a two-fold task, both to decide if a word was a cue phrase or not – a binary task – but also to classify which functional class it belongs to according to an annotation scheme (Hjalmarsson, 2008). The annotators could both see the transcriptions and listen to the recordings while labelling. 81% of the speaker turns con-

tained at least one cue phrase and 21% of all words were labelled as cue phrases. Table 1 presents the distribution of cue phrases over the different classes.

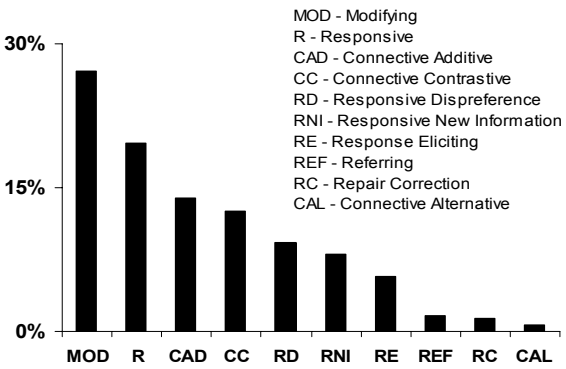

Table 1: Cue phrase distribution over the different classes

The kappa coefficient for the binary task, to classify if a word was a cue phrase or not, was 0.87 (p=0.05). The kappa coefficient for the classification task was 0.82 (p=0.05). Table 2 presents the agreement in percentage distributed over the different classes.

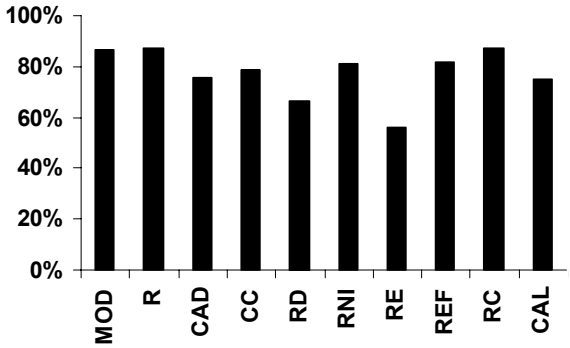

Table 2: % agreement for the different classes

Many cue phrases were used in combinations, signalling function on different discourse levels; first a simple responsive, saying that the previous message was perceived, and then some type of connective to signal how the new contribution relates to previous discourse. The data is a valuable resource of information for how cue phrases are lexically and prosodically realized within in the DEAL domain. To separate cue phrases from other lexical entities and to determine what they signal is a complex task.

3 Analysis of the responsive "ja"

The most frequent lexical item used as a responsive cue phrase in the DEAL corpus was "ja" (Eng: yes). A listening test with 21 subjects, similar to Gravano et al. (2007) was performed to see which is more important for interpretation: the surrounding context or the acoustic realization of the word itself. The listening test contained 23 different stimuli of "ja" taken from the DEAL corpus. All stimuli were presented twice, isolated and in context. The task was to judge which one out of four different discourse-pragmatic categories (Responsive, Responsive New Information, Responsive Dispreference and answer to yes/no question, the sentential meaning of "ja") was most appropriate. Over all, inter-annotator agreements between the two conditions were fairly low. Agreement between the subjects was higher for stimuli in context (k=0.367) than for isolated stimuli (k=0.286), which suggests that contextual cues are important for how "ja" is interpreted. However, a few stimuli received high inter-annotator agreement for both conditions, i.e. with and without context. Acoustic analysis of these stimuli revealed a few characteristic features for some of the different responsive categories. For example: long duration appears as characteristic for the Responsive Dispreference category and the rising shape of the pitch curve, containing two pitch peaks, appears as characteristic for the Responsive New Information Category (for more information on the listening test, see Klarenfjord, 2008).

Acknowledgments

This research was carried out at Centre for Speech Technology, KTH. The research is also supported by the Swedish research council project #2007- 6431, GENDIAL and the Graduate School for Language Technology (GSLT). Many thanks to Rolf Carlson.

References

A. Gravano, S. Benus, H. Chavéz, J. Hirschberg and L. Wilcox. 2007. *On the role of context and prosody in the interpretation of 'okey'*. In proceedings of ACL. Prague, Czech Republic.

A. Hjalmarsson, P. Wik, and J. Brusk. 2007. *Dealing with DEAL: a dialogue system for conversation training.* In Proc. of SigDial. Antwerp, Belgium.

A. Hjalmarsson. 2008. *Speaking without knowing what to say… or when to end.* In Proc. of SigDial. Columbus, Ohio, USA.

J. Klarenfjord. 2008. *Nja, ja and aa: Data collection and analyses of cue phrases with focus on generation of responsives in dialogue systems*. M.Sc. Thesis, KTH, Stockholm.